A sampling approach to measuring a population characteristic of interest may produce a more accurate estimate than a census or complete enumeration. A sample survey requires a much smaller staff than a complete enumeration and it is therefore possible to employ better trained people with higher pay, and also to maintain closer supervision of their work. If the resulting reduction of nonsampling errors is greater than the amount of sampling error that would be introduced, estimates produced from a sample survey will be more accurate than those produced from a complete enumeration. In addition, the sampling approach produces estimates much more quickly and less expensively than a complete enumeration.

Random Sampling Error

In this section we will examine more closely the meaning of sampling error and present some very basic measures of sampling error. The formulas for estimating error that are presented below assume that a **simple random sample** has been drawn from a population and that one wants to estimate characteristics of the entire population based on the results of the single sample. Simple random sampling is the selection from a complete list of a population of a subset of the population so that on any given draw the probabilities of all remaining individuals being selected are equal regardless of the individuals previously selected. For example, assigning the numbers 1 through 1000 to the members of a population of 1000 persons and then randomly selecting 100 different numbers between 1 and 1000 would result in a simple random sample. It should be noted that simple random sampling is but one type of randomized sampling, and that while simple random sampling assumes equal probability of selection, not all equal probability of selection methods are simple random sampling.

Sample designs more complex than simple random sampling are sometimes used and the measures of sampling error associated with these designs are also more complex than the ones presented here. In a **stratified sample** all individuals are first divided into groups or categories and independent samples are then selected within each group or stratum. This method may be used to assure that members of a certain subgroup are included in the sample, or if the strata are relatively homogeneous fewer cases may be required to achieve a given degree of accuracy. In **cluster sampling**, clusters or groups of elements are sampled rather than individual elements.For example, the census tracts in a city might be randomly sampled and then all of the persons in the selected tracts would comprise the sample. Cluster samples yield greater sampling errors than simple random samples of the same size, but the costs associated with cluster sampling are usually considerably less. The general problem is essentially that of balancing cost and efficiency. For large-scale surveys these more complex designs are often desirable, and one should consult a sampling text or sampling specialist for details and assistance (*3*).

A basic concept involved in the determination of sampling error is that of a **sampling distribution**. Consider a population from which we wish to draw a random sample in order to estimate the population mean, $\bar{X}_p$, for example, average number of physician visits per year. Assume that this population has a standard deviation of SD_p, which is a measure of the dispersion of the individual numbers of physician visits around the population average or mean. We would use the mean of the sample, $\bar{X}_s$, to estimate the population mean. If a very large number of random samples were drawn from this population, the distribution of the means of these samples would form the sampling distribution of the mean. There is a "central-limit theorem" in statistics that states: If repeated random samples of size N are drawn from a normally-distributed population, with mean $\bar{X}_p$ and standard deviation SD_p, the sampling distribution of sample means will be normally-distributed, with mean $\bar{X}_p$ and standard deviation $SD_p/\sqrt{N}$. This standard deviation of the sampling distribution is referred to as the **standard error** of the estimate $\bar{X}_s$. This theorem tells us that the larger the sample size selected, the smaller the standard error, i.e., the more the sample means will cluster around the true population mean. Further, to cut the standard error in half we need to quadruple N. Also, the more homogeneous the population is to begin with (the smaller the value of SD_p) the smaller the standard error and thus the greater the clustering of sample means about the population mean. (*4*) Figure 2 portrays the relationship between the population distribution and two possible sampling distributions.

FIGURE 2:

Comparison of Population Distribution and Normal Sampling Distributions for Different Sized Samples